Machine Learning based Drug Indication Prediction using Linked Open Data

Machine Learning based Drug Indication Prediction using Linked Open Data

Remzi Çelebi[1], Özgün Erten[2], and Michel Dumontier[3]

[1] Ege University Computer Engineering Department, Izmir, Turkey,
remzi.celebi@ege.edu.tr,
[2] Ege University Faculty of Medicine, Izmir, Turkey,
ozgun.erten@med.ege.edu.tr,
[3] Institute of Data Science, Maastricht University, Maastricht, Netherlands,
michel.dumontier@maastrichtuniversity.nl

Abstract. In this study, drug and disease features were obtained by querying open linked data to train our classifier for predicting new drug indications, and the predictive performance of the classifier for different validation schemes was evaluated. We collected the drug and disease data from Bio2RDF, an open source project that uses semantic web technologies to link data from multiple sources. A binary feature matrix was generated using drug target, substructure and side effects and disease ontology terms. We collected a broader collection of data containing 816 drugs and 1393 diseases with their features and gold standard data we generated by combining multiple drug indication data sources. We tried our method on a different dataset, compiled by other researchers, that confirmed the predictive value of our method independent of the primary data.
A crucial flaw in the typical evaluation scheme for drug indication predictions that would yield unrealistic predictions is to fail to consider the paired nature of inputs. We partitioned the data in distinct training and test sets where not only pairs but also drugs/diseases are were not overlapped. We tested several classifiers under different cross validation schemes and compared our approach with existing methods. We observed that our model had better predictive performance than the existing models in disjoint cross-validation settings.

Keywords: linked open data, SPARQL, drug repositioning, machine learning, drug indication prediction

1 Introduction

Despite genomic and technological advances, drug discovery and development continues to be a time-consuming and costly process. The number of approved new drugs has remained far below expectations, notwithstanding substantial investments in the pharmaceutical and health sciences. Therefore, one attractive option is to reduce the time and cost of drug development by expanding the scope

of usage of already approved, known drugs. Given that these drugs have passed stringent approvals by US Food and Drug Administration, there is minimal risk associated with their safety and tolerability. Drug repositioning can dramatically reduce development times and costs from the discovery to the clinical approval stages. Between 20 and 30 scientific papers on the subject of drug repositioning are published each month [1]. Given this data, the importance of repositioned drugs on the market is highlighted by the fact that they account for 30% of new indications per year.

Previous efforts to estimate large-scale novel drug indications have focused on the mapping of gene expression profiles [10, 8] and on the recommendation of similar drugs or diseases based on known drug-disease relationships [4]. Machine learning has a significant advantage over other methods, by offering a way in which to optimally combine different drug and disease characteristics into a predictive model. It may also reveal important features that allow for identifying promising drug indications. Machine learning based drug indication prediction studies have used various similarity measures such as chemical structure, side-effect, protein target information. One such approach is the PREDICT method by [5]. In this method, 5 drug-drug similarity and 2 disease-disease similarity measures were used to train a logistic classifier to predict potential drug-disease association. Zhang and colleagues [23] proposed k-nearest neighbor approach (Similarity-based LArge-margin learning of Multiple Sources (SLAMS)) to predict novel drug indications by calculating the combined similarity score with the drug data obtained from different sources. Guney [6] developed an open-source software tool for researchers to repeat this work and made it public.

Since machine learning generally treats drug indication prediction as a binary classification problem, it is necessary to specify the known drug indications (positive set) and the drug-disease pairs with no indications (negative set). Although the indications in the positive set are usually previously known, the results of clinical trials in which drugs have failed are often not reported. A recent attempt aimed to provide a gold standard database, repoDB [1], that also contains failed drug-indications by retrieving clinical trial records from AACT database [4]. But the number of reported failed indications are far less than number of true indications.

In this study, drug and disease features were obtained by querying open data to train our classifier for predicting new drug indications, and the predictive performance of the classifier for different validation schemes was evaluated. We compared our method with previous computational drug indication prediction approaches. We observed that we had better predictive performance than the PREDICT and the SLAMS in disjoint cross-validation settings. Tests and predictions data generated by combining multiple drug indications data sources were evaluated. Finally, we make our work open and freely available so that others can use or extend this methodology [5].

[4] https://www.ctti-clinicaltrials.org/aact-database
[5] https://github.com/rcelebi/drugindication_ml

2 Method

We developed a computational pipeline to reproduce the data and the results of our methodology. The pipeline consists of following steps:

1- Query and download open drug and disease data sets 2- Extract features from data sets 3- Select negative samples and balance the proportion of positive and negative samples that will be introduced into the classifier 4- Apply cross-validation 5- Build classifiers

2.1 Data Compiled from Linked Open Data

Machine learning models to predict drug indications were trained using drug and disease featured extracted from open data.

Most studies use features already curated for drug repurposing. This study generated features obtained from querying repositories of linked data. Linked Data refers to data sources that use Semantic Web technologies to make structured content available on the web. In following the principles of Linked Data, these resources become more FAIR - Findable, Accessible, Interoperable, and Reusable [21]. One key resource for the biomedical sciences on the Semantic Web is Bio2RDF [2], an open source project that uses semantic web technologies to construct and make available a network of linked data from several major biological databases, including Drugbank, KEGG and SIDER. We used Bio2RDF to obtain raw data which were subsequently processed to generate the features for our learning model (Figure 1).

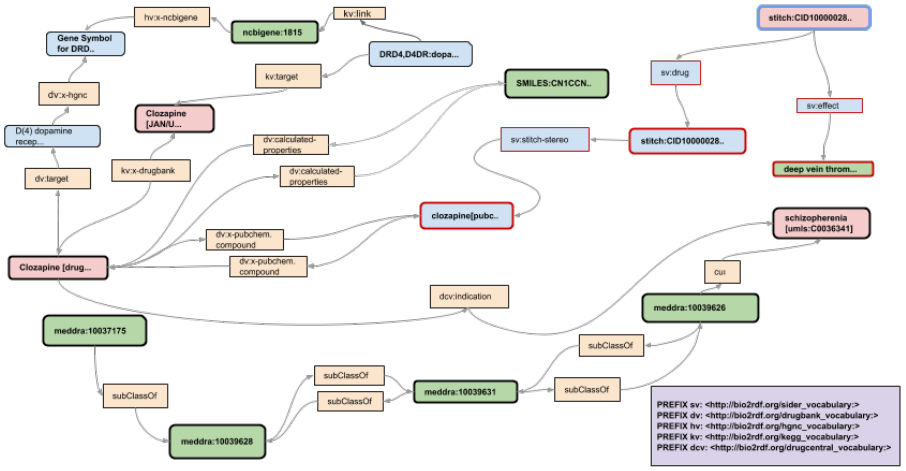

Fig. 1. Visualization of the semantic graph for Clozapine in the subset of Bio2RDF. Bio2RDF normalized and integrated drug data from different data sources in semantic space.

We wrote and executed SPARQL queries to obtain 816 drugs and their targets from DrugBank and KEGG dataset, the chemical structure information of these drugs from DrugBank dataset, the side effect information from SIDER and diseases MedDRA concepts from BioPortal (Noy et al. 2009). In case of a version update to the data, it will be possible to re-execute the queries and obtain new updated data.

We normalized the data using Drugbank identifiers for drugs, NCBI gene identifiers for drug targets and diseases, while side effects were mapped to UMLS identifiers so as to integrate various terminologies.

We obtained drug-disease associations from DrugCentral and The National Drug File Reference Terminology (NDF-RT) repositories. Drug Central contained a total of 6677 drug-disease relationships consisting of 1519 drugs and 1229 diseases. NDF-RT contains 2998 drug indications spanning 782 drugs and 737 diseases that have direct mappings to Drugbank, UMLS concepts respectively. After assembling drug-disease associations, a unified gold standard has 8951 drug indications, 1594 drugs and 1611 diseases (see Table 1). Only 4715 drug-disease associations were used in the experiments where the features could be generated for only 788 drugs and 1103 diseases in the unified gold standard.

Table 1. Statistics about NDF-RT, DrugCentral and unification of two gold standards.

	DrugCentral	NDF-RT	Common	Unified
Drug	1519	782	707	1594
Disease	1229	737	355	1611
Drug-Disease Association	6677	2998	724	8951

2.2 Extracting Features

Chemical structure Drug structure at the molecular level describes its binding activity. Chemical fingerprints are the most commonly used structural profiling marker for drugs [13]. Fingerprints are bit vectors that indicate the presence (1) or absence (0) of certain chemical features (e.g. a C=N group, a six member ring,). We used the OpenBabel 2.3 library to take an input chemical formula (SMILES ID) and generate Molecular Access System (MACCS) binary structural feature lists with lengths of 166.

Drug targets The set of targets for a drug can shed light on affected biological processes. We represent the set of drug targets obtained from DrugBank and KEGG as a bit vector in which 1 represents a target of the drug, and a 0 represents not a target for the drug. This results in a sparse matrix, since the drugs have a median of one putative target each.

Drug Side Effects Side effects elicited by drugs are suggestive of a physiological role. Previous studies have used used side effects to estimate drug similarity, despite the potential noise in labeling [3, 22]. We used SIDER [9] as a source of drug side-effect information. SIDER was automatically constructed by text mining of drug product labels and are known to contain false positives.

Disease Description A drug can be indicated for a greater number of diseases than its original indications. In order to be gain information about these situations, it is necessary to produce profiles that describe the level of similarity between diseases. In order to produce a disease profile, we used top-level concepts that the disease shared on an ontology. We obtained NDF-RT and MedDRA ontologies from BioPortal to define a disease with its top-level concepts. If a disease is present in a ontology, the top concepts associated with this concept represent 1 (existence) or 0 (absence) in the feature vector.

2.3 Selecting and Balancing Negative Samples

We tested the strategy for the selection of negative examples that was conducted by means of random selection of negative cases from among unknown drug-disease associations within the diseases at least one drug indicated for. The negative set is randomly selected from unknown drug-disease associations in some proportion to the number of pairs within the the positive set. The user can input the proportion of the positive and negative samples within each fold.

2.4 Evaluation

Existing studies generally predict that the drugs in the test set will also be in the training set. However, researchers are more interested in discovering a drug whose indications are unknown, so the evaluation established in this way can give misleading information about the prediction of indications for new drugs. Guney [6] examined the situation where the drugs in the test set are disjoint from the drugs in training set. We have expanded Guney's drug-wise cross-validation approach to include disease-wise cross-validation as well (see Figure 2). Thus, prediction performance changes were observed in the samples in the test set differed from those in the training set, in which they have no common drugs or diseases. We used these different cross-validation schemes to see how reliable our estimates are for a drug or disease that is not in the training set.

2.5 Building Classifiers

We used Python Scikit-learn machine learning package to build the classifiers. Various classifiers were constructed with logistic regression (LR), k-nearest neighbor classifier (KNN), random forest (RF), and gradient boosting classifier (GBC). The parameters for building different classifiers were chosen as follows: L2 penalty and $C = 1.0$ for LR; $n\_neighbors = 5$ for KNN; $n\_estimators = 1000$ and $max\_depth = 5$ for RF and GBC. We implemented approaches for data balancing, cross validation and classifier building.

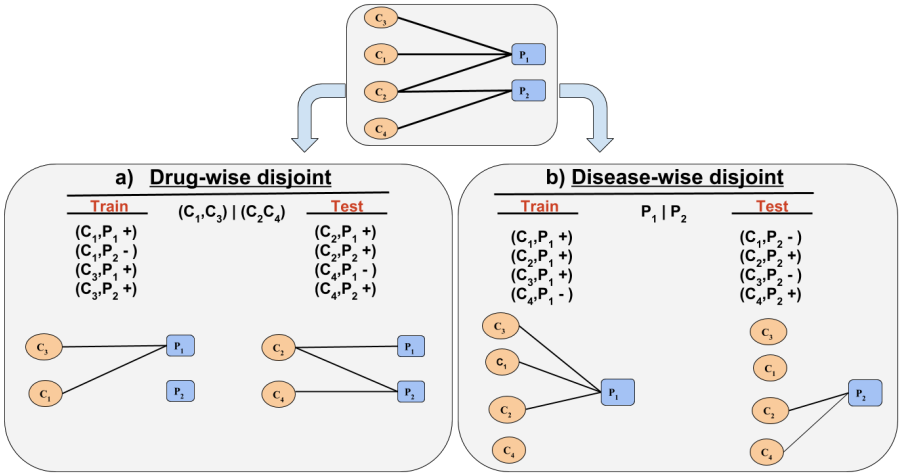

Fig. 2. Graphical representation of training-test split of a toy data. In this example c1, c2, c3, and c4 correspond to four compounds, and p1 and p2 correspond to two phenotypes. A known drug indication between a compound and phenotype is represented as the edge of the graph. Since two-fold cross-validation was used in this example, two groups were separated in terms of drug or disease. In the case of a) drug and b) disease and related associations are grouped into training and test data.

3 Results

We first compared our approach with the SLAMS method using NDF-RT gold standard and the data already curated, available online through Guneys tool. By trying the same data used in the SLAMS method we wanted to show the predictability of our method independent of the data compiled. The NDF-RT version that they used contained a total of 3250 drug relationships between 799 drugs and 719 diseases. We observed best AUC = 87.10% for the NDFRT gold standard using Gradient Boosting Tree Classifier with pair-wise cross-validation (see Table 2). AUC fell to 82.77% in drug-wise cross-validation (no two drugs are not in the same fold) . Here, the number of negatives samples chosen was twice as large as the positive set. In comparison, the SLAMS could yield best AUC =84.65% with Logistic Regression for in pair-wise cross-validation and AUC =68.43% in drug-wise cross-validation. It shows us there is a huge improvement in prediction performance for drug-wise (68.43% to 82.77%) and pair-wise (84.65% to 87.10%).

We next examined the prediction performance of our method with the unified indication gold standard and the data compiled from open linked data. Figure 3 shows the AUC for different classifiers under various validation schemes averaged over ten runs of ten-fold cross validation. We observed Gradient Boosting Classifier (GBC) has significant prediction performance over other ML methods with AUC of 0.88. Under both drug-wise and disease-wise cross validation schemes,

Table 2. Areas under ROC curves (AUC) under drug-wise and pair-wise cross-validation averaged over ten runs of ten-fold cross validation for NDF-RT gold standard.

		Our Method	SLAMS
Model	Drug disjoint	AUC	AUC
LR	no	73.03 ± 2.00	84.65 ± 0.19
	yes	69.46 ± 3.79	68.43 ± 0.87
RF	no	82.88 ± 1.69	82.78 ± 0.40
	yes	76.79 ± 3.78	65.27 ± 0.82
KNN	no	70.81 ± 2.26	81.83 ± 0.82
	yes	70.52 ± 4.22	65.44 ± 0.75
GB	no	87.10 ± 1.53	84.22 ± 0.36
	yes	82.77 ± 3.37	67.82 ± 0.74

GBC was better than other ML algorithms and did not fall below AUC score of 0.83.

When considering drug-wise disjoint cross-validation scheme, SLAMS obtains AUC score of 0.66 with logistic regression at best. Another observation is PREDICT with the drug and disease similarities using the same data obtains an averaged AUC score of 0.72 with logistic regression under the same scheme.

3.1 Analysis of Novel Predictions

To evaluate the predictive power of our method, we investigate the predictions made by our tool for drug Reboxetine. Reboxetine is an antidepressant effective drug in the selective noradrenaline reuptake inhibitor (SNARI) group used in the treatment of depression with high affinity for the carrier of noradrenaline, which selectively inhibits noradrenaline reuptake in the presynaptic range.

Reboxetine has only one indication (Major Depressive Disorder) specified in our gold standard. In the light of current literature, Reboxetine is also suggested as an effective and safe option for the treatment of depression, sleep disorders [11, 17], eating disorders [7, 18], attention deficit hyperactivity disorder (ADHD) [19, 16], panic attack [20], depression in parkinsonian patients [12]. The estimates we made for potential indications of this drug are given in the Table 3.

The probabilities for the potential indications for the logistic classifier Reboxetine were given in Table 3 and the average probability of 17 indications are 0.65. For the first 15 diseases, the probability is greater than 0.5 and it is understood that the indication is likely. Our model predicts that among all diseases, 200 diseases may be associated with Reboxetine ($P > 0.5$). In addition to the reasonable estimates such as Hypertensive disease ($P = 0.937$) and Allergic rhinitis ($P = 0.986$), which need to supported by evidence from the literature.

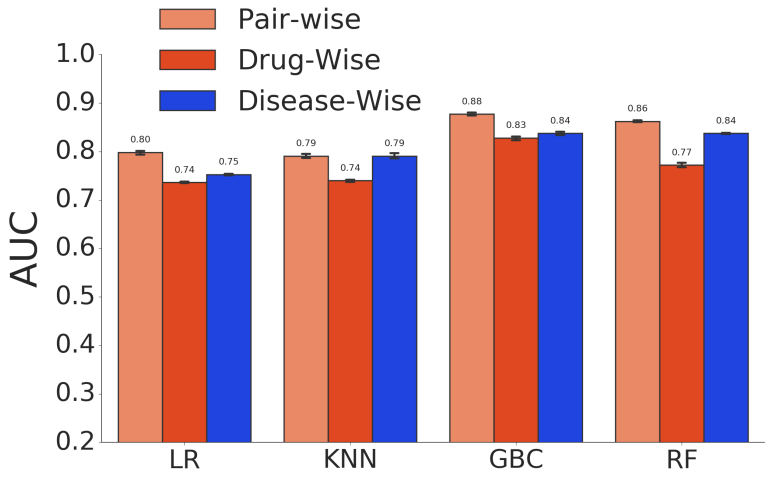

Fig. 3. Areas under ROC (AUC) under various validation schemes averaged over ten runs of ten-fold cross validation for the unified gold standard.

4 Conclusion

Researchers have exploited publicly accessible datasets to validate their hypotheses for prediction of drug indications. However, the datasets are diverse and are subject to change over time, which may result in different conclusions for the same hypotheses. We used Semantic Web technologies, specifically Linked Data, to represent, link and access data related to drugs and diseases provided by the Bio2RDF project. We use SPARQL queries to obtain drug and disease features to train classifiers. In case of a version update to the data, it will be possible to re-execute the queries and obtain new updated data.

We collected a wider collection of data containing 816 drugs and 1393 diseases with their features. Predictions for gold standard data generated by combining multiple drug indications data sources were evaluated. We tried our method on a different dataset, compiled by [23], that show us the predictability of our method independent of the data compiled.

A crucial flaw in a typical evaluation scheme for drug indication predictions that would make unrealistic predictions is failure to consider the paired nature of inputs [15]. We partitioned the data in distinct train and test sets where not only pairs but also drugs/diseases are not overlapped as suggested in [14] for drug-target interaction prediction. We tested several classifiers under different cross validation schemes and compared our approach with existing methods namely PREDICT, SLAMS. We observed that we had better predictive performance than the PREDICT and the SLAMS in disjoint cross-validation settings.

Table 3. Potential indications for Reboxetine and prediction scores by our model

Ranking	Disease	Probability
1	Narcolepsy	0.82
2	Depressive disorder	0.8
3	Parkinson Disease	0.78
4	Schizophrenia	0.77
5	Obsessive-Compulsive Disorder	0.71
6	Anxiety Disorders	0.71
7	Generalized Anxiety Disorder	0.71
8	Panic Disorder	0.69
9	Obesity	0.69
10	Cerebrovascular accident	0.67
11	Anorexia nervosa	0.67
12	Bulimia Nervosa	0.67
13	Attention deficit hyperactivity disorder	0.63
14	Eating Disorders	0.63
15	Post-Traumatic Stress Disorder	0.58
16	Fibromyalgia	0.46
17	Binge eating disorder	0.18

Acknowledgement. The first named author (R.C.) is grateful to TUBITAK for providing financial support under 2214-A programme.

References

1. Brown, A.S., Patel, C.J.: A standard database for drug repositioning. Scientific Data 4, 170029 (2017)
2. Callahan, A., Cruz-Toledo, J., Ansell, P., Dumontier, M.: Bio2rdf release 2: improved coverage, interoperability and provenance of life science linked data. In: Extended Semantic Web Conference. pp. 200–212. Springer (2013)
3. Campillos, M., Kuhn, M., Gavin, A.C., Jensen, L.J., Bork, P.: Drug target identification using side-effect similarity. Science 321(5886), 263–266 (2008)
4. Chiang, A.P., Butte, A.J.: Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. Clinical Pharmacology & Therapeutics 86(5), 507–510 (2009)
5. Gottlieb, A., Stein, G.Y., Ruppin, E., Sharan, R.: Predict: a method for inferring novel drug indications with application to personalized medicine. Molecular systems biology 7(1), 496 (2011)
6. Guney, E.: Reproducible drug repurposing: When similarity does not suffice. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017. pp. 132–143 (2017)
7. Hay, P.J., Claudino, A.M.: Bulimia nervosa: online interventions. BMJ clinical evidence 2015 (2015)
8. Hu, G., Agarwal, P.: Human disease-drug network based on genomic expression profiles. PloS one 4(8), e6536 (2009)
9. Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The sider database of drugs and side effects. Nucleic acids research 44(D1), D1075–D1079 (2015)

10. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., et al.: The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. science 313(5795), 1929–1935 (2006)
11. Larrosa, O., de la Llave, Y., Barrio, S., Granizo, J.J., Garcia-Borreguero, D.: Stimulant and anticataplectic effects of reboxetine in patients with narcolepsy: a pilot study. Sleep 24(3), 282–285 (2001)
12. Lemke, M.R.: Effect of reboxetine on depression in parkinson's disease patients. The Journal of clinical psychiatry 63(4), 300–304 (2002)
13. Melville, J.L., Hirst, J.D.: TMACC: Interpretable Correlation Descriptors for Quantitative StructureActivity Relationships. J. Chem. Inf. Model. 47(2), 626–634 (Mar 2007), http://dx.doi.org/10.1021/ci6004178
14. Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szwajda, A., Tang, J., Aittokallio, T.: Toward more realistic drug–target interaction predictions. Briefings in bioinformatics 16(2), 325–337 (2014)
15. Park, Y., Marcotte, E.M.: Flaws in evaluation schemes for pair-input computational predictions. Nature methods 9(12), 1134–1136 (2012)
16. Ratner, S., Laor, N., Bronstein, Y., Weizman, A., Toren, P.: Six-week open-label reboxetine treatment in children and adolescents with attention-deficit/hyperactivity disorder. Journal of the American Academy of Child & Adolescent Psychiatry 44(5), 428–433 (2005)
17. Schmidt, C., Leibiger, J., Fendt, M.: The norepinephrine reuptake inhibitor reboxetine is more potent in treating murine narcoleptic episodes than the serotonin reuptake inhibitor escitalopram. Behavioural brain research 308, 205–210 (2016)
18. Silveira, R.O., Zanatto, V., Appolinario, J., Kapczinski, F.: An open trial of reboxetine in obese patients with binge eating disorder. Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity 10(4), e93–e96 (2005)
19. Tehrani-Doost, M., Moallemi, S., Shahrivar, Z.: An open-label trial of reboxetine in children and adolescents with attention-deficit/hyperactivity disorder. Journal of child and adolescent psychopharmacology 18(2), 179–184 (2008)
20. Versiani, M., Cassano, G., Perugi, G., Benedetti, A., Mastalli, L., Nardi, A., Savino, M.: Reboxetine, a selective norepinephrine reuptake inhibitor, is an effective and well-tolerated treatment for panic disorder. The Journal of clinical psychiatry (2002)
21. Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., da Silva Santos, L., Bourne, P., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R., Gonzalez-Beltran, A., Gray, A., Groth, P., Goble, C., Grethe, J., Heringa, J., 't Hoen, P., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S., Martone, M., Mons, A., Packer, A., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M., Thompson, M., Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The fair guiding principles for scientific data management and stewardship. Scientific Data 3 (2016)
22. Yang, L., Agarwal, P.: Systematic drug repositioning based on clinical side-effects. PloS one 6(12), e28025 (2011)
23. Zhang, P., Agarwal, P., Obradovic, Z.: Computational drug repositioning by ranking and integrating multiple data sources. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 579–594. Springer (2013)